Incumbency Advantage Using Data on U.S. House Election Returns

Selina Wu

February 12, 2020

```
setwd("C:/Users/selin/OneDrive/Documents/Data Projects/PoliSci/1_Attempt 2")

h <- read.csv("house.csv")

dem.pct <- 100*((h$vote_D)/(h$vote_D+h$vote_R))
h$dem_voteShare <- dem.pct
```

The house data is stored in the variable "h". Democratic vote percentage is computed for each election and stored in the data frame as a new column, titled "dem_VoteShare".

Mean Democratic vote percentage

The mean function is used to calculate the mean of all values stored in dem.pct, with missing data removed.

```
mean_ov <- mean(dem.pct, na.rm = T)
mean_ov

## [1] 56.12524
```

Mean Democratic vote percentage when Democratic incumbent is running

```
mean_Dem <- mean(h$dem_voteShare[h$inc_D>=1], na.rm=T)
mean_Dem

## [1] 72.9761
```

The mean function is used to calculate the mean Democratic vote percentage when a Democratic incumbent is running. Only the rows in which the inc_D value is equal to 1 or greater will be looked at under this condition. The dem_voteShare column values for these specified rows are fed into the mean function as the argument, with missing data removed.

Mean Democratic vote percentage when NO Democratic incumbent is running

```
mean_noDem <- mean(h$dem_voteShare[h$inc_D==0], na.rm=T)
mean_noDem

## [1] 41.32496
```

The mean function is used to calculate the mean Democratic vote percentage when a Democratic incumbent is not running. Only the rows with inc_D value equal to 0 will be looked at under this condition. The dem_voteShare column values for these specified rows are fed into the mean function as the argument, with missing data removed.

Mean Democratic vote percetage when neither Democratic nor Republican Incumbent is running

```
mean_open <- mean(h$dem_voteShare[h$inc_D==0 & h$inc_R==0], na.rm=T)
mean_open

## [1] 51.83472
```

The mean function is used to calculate the mean Democratic vote percentage when neither a Democratic nor Republican incumbent is running. Only the rows with inc_D and inc_R values equal to 0 will be looked at under this condition. The dem_voteShare column values for these specified rows are fed into the mean function as the argument, with any rows with missing data removed.

Comparing the three Democratic vote percentages

The mean Democratic vote percentage (DVP) is highest at 72.98% when a Democratic incumbent is running, and lowest at 41.32% when no Democratic incumbent is running, but a Republican incumbent can be running. When neither party has an incumbent running, the Democratic vote percentage is closer to the average overall Democratic vote percentage, 51.83%, and to 50%.

Since the DVP is highest when a Democratic incumbent is running, lowest when no democratic incumbent is winning, and higher than when neither party has an incumbent running, this data suggests that an incumbent running for election often receives more votes and performs better than when a candidate who is not an incumbent runs. Additionally, since the DVP is lowest when no Democratic incumbents – but some Republican incumbents – run, this suggests that incumbents from each party are more likely to gain more votes for their respective party.

Interpretation between mean Democratic vote share when Democratic incumbent running and that when neither party has an incumbent running

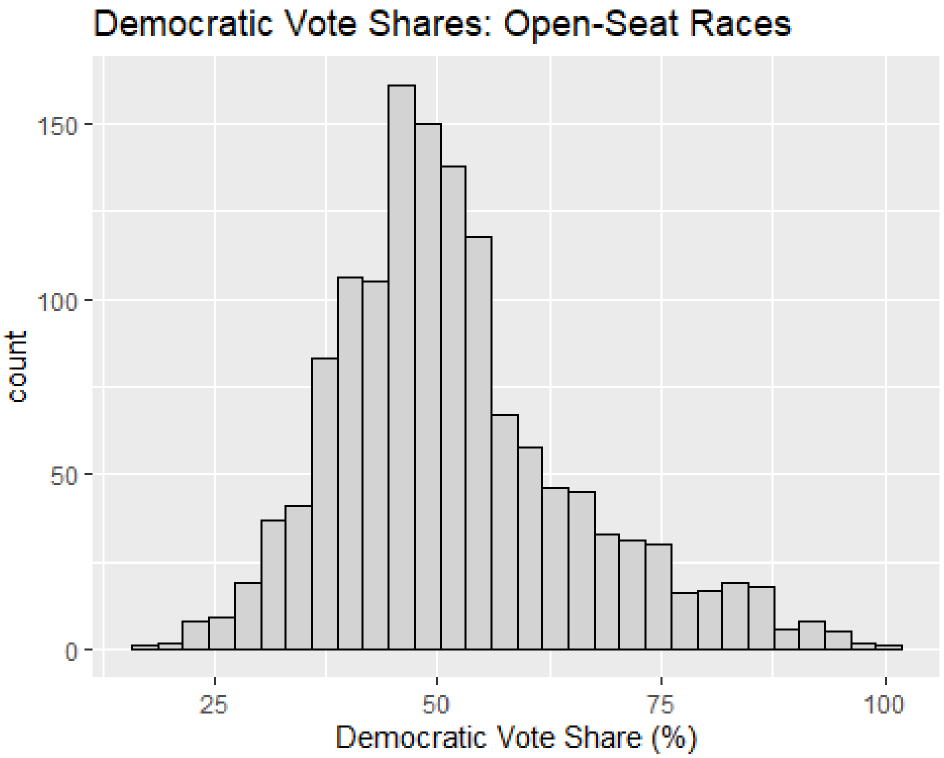
```
diff.dpct <- mean_Dem - mean_open
diff.dpct

## [1] 21.14138
```

The difference between the mean Democratic vote share in elections with a Democratic incumbent and that of open-seat races is 21.14%. It is difficult to assume causal inference because we can never observe the same election as if the incumbent who ran did not run, and vice versa. As such, I do not think this difference of 21.14% is causal. Voting behavior is not inherently randomized because people have non-random ideologies, and the 'treatment' of having a Democrat run is not randomly assigned. The election and people's votes can vary in many aspects between when an incumbent runs or does not run, so it is extremely difficult to assume similarities in the election for if an incumbent runs/does not run. For these reasons, I believe causality cannot be concluded.

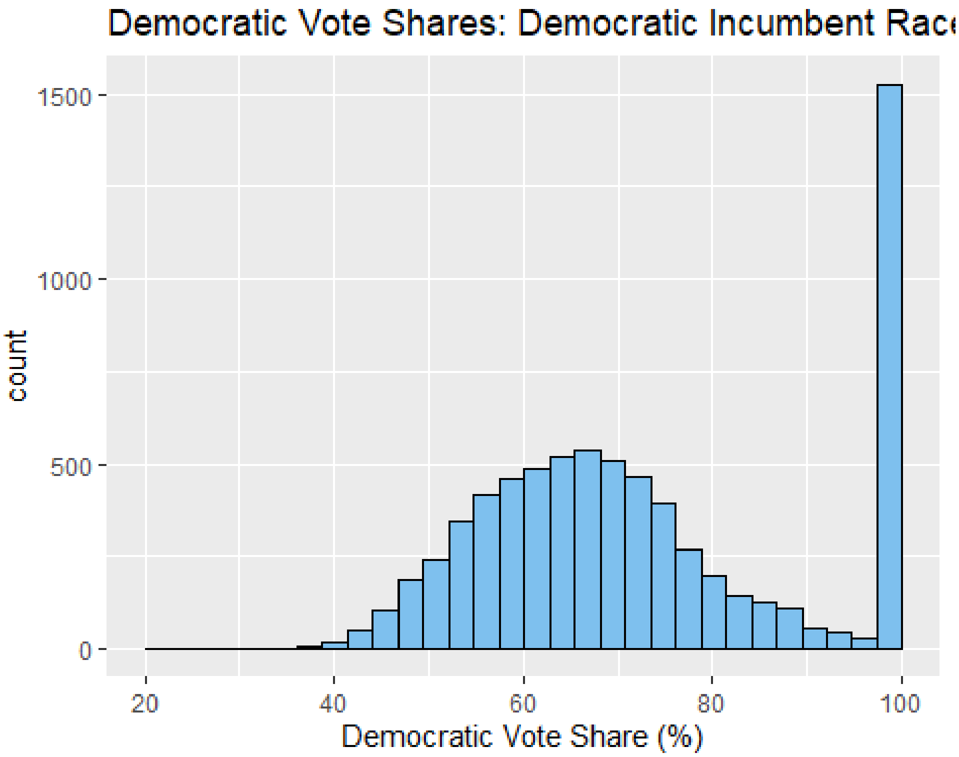
Visual of DVS for races with a Democratic incumbent and open-seat races

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3

df.OpenSeat <- h[(h$inc_D==0) & (h$inc_R==0), c(4:8)]

ggplot(df.OpenSeat, aes(x=dem_voteShare))+geom_histogram(color="black", fill=
"light gray")+
  labs(title="Democratic Vote Shares: Open-Seat Races", x="Democratic Vote Sh
are (%)")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 87 rows containing non-finite values (stat_bin).
```



```
df.DemInc <- h[(h$inc_D==1) | (h$inc_D==2), c(4:8)]
ggplot(df.DemInc, aes(x=dem_voteShare))+geom_histogram(color="black", fill="s
kyblue2")+
  labs(title="Democratic Vote Shares: Democratic Incumbent Races", x="Democra
tic Vote Share (%)")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 459 rows containing non-finite values (stat_bin).
```



ggplot was used to plot histograms for Democratic vote shares for open-seat and Democratic incumbent races, respectively.

Only rows in which inc_D and inc_R values equal 0 and columns 4-8 were selected and subsetted into a new data frame, stored as "df.OpenSeat". The "dem_voteShare" column values were plotted on the x-axis, and the frequency ("count") on the y-axis.

From the house data frame, only rows in which inc_D equal 1 or 2 and columns 4-8 were selected and subsetted into a new data frame, stored as "df.DemInc". The "dem_voteShare" column values were plotted on the x-axis, and the frequency ("count") on the y-axis.

Interpretation of histogram

In the "Democratic Vote Shares: Democratic Incumbent Races" histogram, there is lumping at the 100% values. This occurs because there are races in which a Democratic incumbent is running while there is missing data for whether or not a Republican incumbent is running. When this occurs, the number of votes for the Republican candidate equals 0, so the Democratic vote percentage for these races is 100%. According to the histogram, there are over 1500 races for which Democratic vote percentage is 100%.

The other histogram does not have this same lumping because the data considered in calculating the Democratic vote shares for these races only includes rows in which both inc_D and inc_R values equal 0. For these races, all candidates – Democrat or Republican – received some votes. The number of elections in which Democratic vote percentage is ~99-100% in open-seat races is minimal.

Average Democratic vote share by year for all election years

```
avg.DVS_yr <- aggregate(h$dem_voteShare, list(h$year), mean, na.rm=T)
colnames(avg.DVS_yr) <- c("Year", "Mean_DVS")
avg.DVS_yr
```

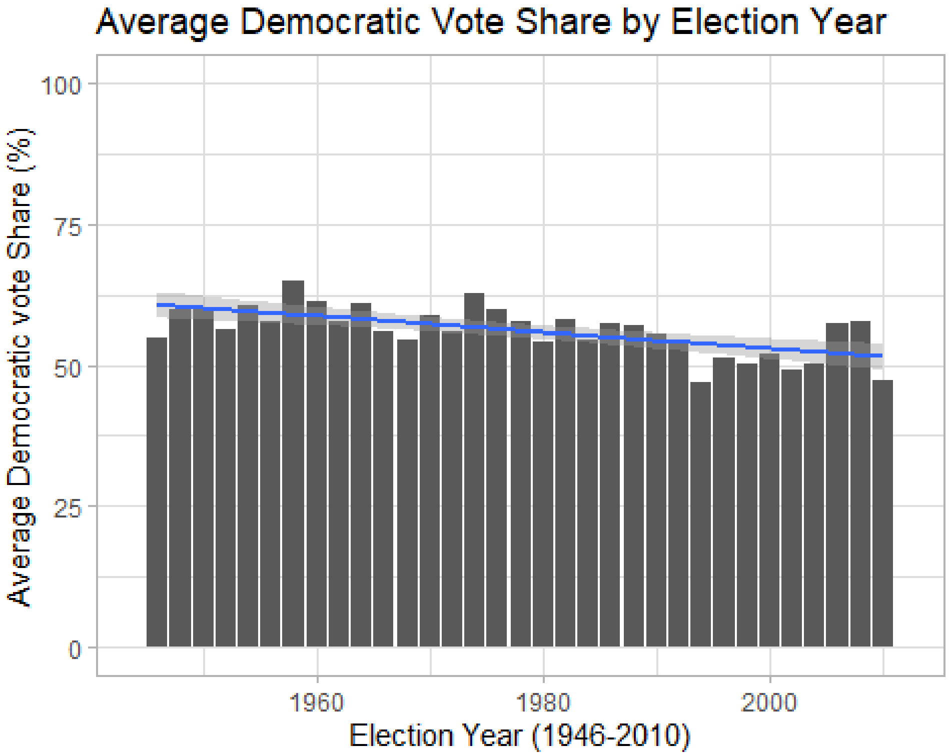
```
##    Year Mean_DVS
## 1  1946 54.78213
## 2  1948 59.84499
## 3  1950 59.80617
## 4  1952 56.43894
## 5  1954 60.60806
## 6  1956 57.65365
## 7  1958 64.83380
## 8  1960 61.24619
## 9  1962 57.87921
## 10 1964 61.03573
## 11 1966 55.89058
## 12 1968 54.47377
## 13 1970 58.85662
## 14 1972 56.00020
## 15 1974 62.87663
## 16 1976 59.88538
## 17 1978 57.55585
## 18 1980 54.01052
## 19 1982 57.90683
## 20 1984 54.67160
## 21 1986 57.39859
## 22 1988 56.93204
## 23 1990 55.53235
## 24 1992 53.92405
## 25 1994 47.06497
## 26 1996 51.22967
## 27 1998 50.02364
## 28 2000 51.90223
## 29 2002 49.02841
## 30 2004 50.26722
## 31 2006 57.20717
## 32 2008 57.70137
## 33 2010 47.28203
```

An aggregate function was used to collate dem_voteShare values from the house data frame, and this data was grouped by year. The mean function is the function applied to the dem_voteShare column, with missing data removed for the calculation.

Over the 33 years, the average Democratic vote share appears fairly even in the 50% range, though seems to be decreasing very slightly since around 1960. With the exception of some elections having mean Democratic vote percentages in the 40%'s and 60%'s, the values remain in the 50%s. This suggests that Democratic and Republican vote shares are fairly level with each other. However, there appear to be more mean Democratic vote percentage values over 50% and less under 50%, suggesting Democrats have mostly won more votes overall since the 1946 election.

```r
ggplot(avg.DVS_yr, aes(y=Mean_DVS, x=(Year)))+geom_bar(stat="identity")+theme_light()+
  stat_smooth(method="lm")+
  ggtitle("Average Democratic Vote Share by Election Year")+
  xlab("Election Year (1946-2010)")+
  ylab("Average Democratic vote Share (%)")+
  scale_y_continuous(limits=c(0, 100)) + scale_x_continuous(limits=c(1944, 2012))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



ggplot was used to plot a bar plot the results of Question 11.

A bar plot was used since the election years are discrete, and the years are plotted along the x-axis. Each bar represents an election year, starting from 1946 increasing by 2 years until 2010. The average Democratic vote share for that year is plotted along the y-axis, and ranges from 0-100%. A trendline was added to show the Democratic vote share slightly decreasing overall in these years, as the slope of the trendline is negative.

The bars leveling around 50% shows how Democratic and Republican vote shares are relatively level with each other, as the Republican vote shares would be from the top of the bars to 100%. The bar plot also shows the election years when average Democratic vote share was noticeably higher or lower than that of the surrounding years.

Average Democratic vote share for any given state

```r
state_dpct <- function(a,b,c){
  avg.DVS <- aggregate(b, list(a), mean, na.rm=T)
  colnames(avg.DVS) <- c("State", "Mean_DVS")

  avg_byState <- avg.DVS[avg.DVS$State==c,]
  return(avg_byState)
}

state_dpct(h$state, h$dem_voteShare,"MA")
```

```
##    State Mean_DVS
## 19    MA 67.36723
```

```r
state_dpct(h$state, h$dem_voteShare,"TX")
```

```
##    State Mean_DVS
## 43    TX 67.51765
```

A function was created which takes three arguments: a is a placeholder for the column of states, b is a placeholder for the vote shares, and c is a placeholder for the name of the state which will be subsetted on.

I expected the average Democratic vote share for Massachusetts to be relatively high and over 50%, which it is. This is because I know Massachusetts is a Democratic State and usually votes for the Democratic candidate.

On the other hand, I did not expect the average Democratic vote share for Texas to be so high – especially not higher than that of MA. I have known Texas to be a Republican state, so I expected the average Democratic vote share to be under 50%. When I revisited the data, I saw there was much missing data for Republican votes, so many TX elections were calculated to have Democratic vote Shares of 100%.